AN AUTOMATIC CORRELATION METHOD FOR GENERATING SUMMARIES FOR TEXT DOCUMENTS

BACKGROUND OF THE INVENTION

Field of the Invention

This invention relates to an automatic text processing method, and in particular, to a method for generating summaries for text documents.

Background Description

In information query, for the user's convenience, it is normally required that the summaries are generated for the users by means of automatic text processing functions of the computers. The current practical methods for automatically generating summaries for text documents are following four kinds:

- List the first paragraph of the document or the beginning paragraphs of the document as the summary (Infoseek, Yahoo, etc.): This method is simple, but does not suit for common document style;
- List the sentences in the query commands (Lotus website, Beijing Daily Online, etc.): The listed sentences relate directly to the query, but cannot represent the overall style of the document;
- Use implicit template: This method matches some patterns in the document, and then fills the matched contents into the pre-formed template. This method can generate very fluent summaries, but can only be suitable for a fixed document style and a specific domain, and is very difficult to be used commonly;
- Count the occurrence frequency of words or characters: This is a statistics-based method, which generally can be divide into four steps: (1) analyze the document discourse, and segment the document into paragraphs and sentences; (2) segment the sentences into words; (3) evaluate the scores of the importance of the words and the

sentences; (4) output the sentences with higher evaluated scores as the document's summary.

Although the above statistics-based method for automatically generating summaries for text documents has considered the occurrence frequency of words and characters in documents and therefore evaluated the importance of the words and the sentences, the summaries can not well correspond to the user's requirements because there is no interaction with the user. Therefore, the invention proposes a method for automatically generating summaries for text documents, which, when receiving the user's text documents, queries the fields, topics, and terms that the user is interested in. The method extracts the important sentences, and then in reasonable order, outputs them as the document's summary. The method can not only generate summaries for respective documents, but also generate a comprehensive prompt for the important ideas of the documents.

SUMMARY OF THE INVENTION

The method for automatically generating summaries for text documents according to the invention includes steps of:

For a set of documents, generating a set of sentences by document discourse analysis, and obtaining a set of words by morphologic processing;

Initializing a score for each word in the set of words, and each sentence in the set of sentences;

Computing the score for each word in the set of sentences according to the scores of sentences containing it and the correlation degree between the word and the user information;

Computing the score for each sentence in the set of sentences according to the scores of words composing it and the position of the sentence in a section and a paragraph;

If the sum of scores of the words and the sum of scores the sentences change apparently, go back to the step of computing the word scores, otherwise continuing.

Outputting the top-ranked sentences as the summary of the set of documents, the top-ranked words as keywords list of the set of documents.

The above method introduces the following functions into the traditional statistics-based methods:

- It has a new sentence ranking strategy called "automatic correlation method", which is much responsive to user's requirements;
- It supports user summarization profile, which allows a user to customize the fields, topics, and terms that the user is interested in.
- It applies to general-purpose, and is also suited for summarizing the certain query documents.

The method considers the following factors when computing the scores of words in a word set: the correlation degree between the word and the user's summarization profile language; the similarity degree between the word and the query term or topic provided by the user; sum scores of the sentences to which the word belongs; the similarity degree between the word and the word terms in the document title; the ratio of the occurrence times of the word in the document to its occurrence times in the document set; and the ratio of number of the documents in which the word occurs, to the total number of the documents contained in the document set.

BRIEF DESCRIPTION OF THE DRAWINGS

The advantages and features of the invention will be more apparent by the description of preferred embodiments of the invention in conjunction with the accompanying figures, in which:

Fig. 1 is a flow chart of the method for automatically generating summaries for text documents according to a particular embodiment of the invention; and

Fig. 2 is a flow chart used in the sentence ranking part in Fig. 1.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

As shown in Fig. 1, a method for automatically generating summaries for text documents according to a particular embodiment of the invention includes steps of:

Step 1. Document Discourse Analysis

This step identifies titles, sections, lists, paragraph boundaries and sentence boundaries of the documents.

Step 2. Morphologic Process

This step performs morphologic process for each sentence according to the language of documents. For Chinese language, the morphologic process includes the steps of: (1) segmenting sentences into words according to the system dictionary and the user-defined dictionaries; (2) identifying proper names (currently including person names, place names and person titles), domain terms, numbers, measure words, and date expressions; (3) adding part-of-speech tags to each word; (4) resolving the pronominal anaphora of the personal pronouns; (5) identifying the word relationship (such as same object names, synonyms, concept relationships, etc.) and building a relationship network among all of the words, according to the system thesaurus and user thesaurus. For English language, this step includes the steps of: (1) normalizing terms to a standard term; (2) identifying proper names; (3) splitting compound terms; and (4) filtering stop-word.

Step 3. Sentence Ranking

This step is to determine the importance of eachsentence by an automatic correlation algorithm. This step is described in more details below.

Step 4. Summary Output:

- If the user requires one summary for one document, then this step selects the top-ranked sentences to output according to their appearing order in this document;
- If the user requires to a single comprehensive summary for the document set, then this step outputs the sentences according to their computed scores from high to low and marks the sentences to show from which document they come (for example, by adding hyperlinks to the sentences), so that the user can easily look up the respective document.

In both cases the pronouns will be replaced by their entities.

After the document discourse analysis and the morphologic process are performed for each document in the document set D, each sentence in the document set is ranked to decide its important degree according to the sentence set S and the word set W of each document. The sentences are ranked to compute their scores from each other by using an

autonomous correlation method, i.e. by using the sentence set S and the word set W. This is realized by the steps below (see Fig. 2):

Step 1. Introducing a data group SCORE to record the computed scores of the sentences and the words, and initializing a score SCORE[s] of each sentence and a score SCORE[w] of each word into 0;

Step 2. Computing a score SCORE[w] of each word according to the weighted-average of the following values:

The correlation degree between W and the user's summarization profile language;

The similarity degree between W and the query terms or topic provided by the user;

Sum scores of the sentences to which W belongs;

The similarity degree between W and the word terms in each document title;

The ratio of the occurrence times of W in the document to its occurrence times in the document set; and

The ratio of number of the documents in which W occurs, to the total number of the documents contained in the document set D;

This can be written by the following formula, i.e.

$$\begin{aligned}SCORE[w]=&\lambda_1*salience(w, \text{user summarization profile})\\&+\lambda_2*salience(w, \text{user's query or topic})\\&+\lambda_3*\Sigma(SCORE[s], s\ni\omega)\\&+\lambda_4*salience(w, \text{tile words})\\&+\lambda_5*FREQUENCY(w/d)/FREQUENCY(w/D)\\&+\lambda_6*NUMBER(d, d\ni w)/\ NUMBER(D)\end{aligned}$$

Formula 1

Step 3. Computing the SCORE[s] of the sentence according to the weighted-average of the three values below:

- Sum scores of all the words constituting the sentences;
- The position of the sentence in the paragraph and section: the fist sentences in the paragraph and section get higher scores than the sentences in other positions;
- The similarity among the sentences: if in many documents there are sentences whose contents are similar, these sentences are weighted more;

This can be written as the following formula

$$SCORE[s]=\lambda_7*\sum(SCORE[w], s\ni w)+\lambda_8*position(s, d)+\lambda_9*similarity(s, S)$$

Formula 2

Step 4. If the sum of the all scores changes significantly, cycle Step 2; otherwise the process ends.

It is seen according to the description of the invention in conjunction with the particular embodiments that the summary method of the invention is also a statistics-based method, which performs the discourse analysis and the morphologic process, and that the new functions of the method are:

- Allowing the user to customize "the user summary profile" in which the user can list the fields and topics he or she interested in, and also listed he or she is sensitive to which kind of word (such as person names, person titles, place names, numbers, amounts of money, dates, terms defined by the user own, etc.);
- The generated summaries being capable of being driven by subjects or the user's query;
- A completely new sentence ranking strategy, herein called "automatic correlation method": the first step, initializing the ranking score for the words and sentences; the second step, computing scores for every words according to the user summarization profile, topics or query terms provided by the user, and frequencies of the words; the third step, computing the ranking scores according to the words contained in the sentences and the document discourse in the document set; the fourth step, feedbacking the sentence scores to the words and repeating the second step and the third step, until the sentence scores have been stabilized.

This method fully utilizes the dicourse information of every document, the clue words in the document, the title words, the language processing results, and topics or query terms provided by the user, to make the generated summaries accord with the user's requirements more completely.

The flow charts described here are only exemplary, and many modifications can be made to those chart examples or the steps (or operations) described therein without departing from the spirit of the invention. For example, those steps can be executed in different order, or increased, reduced or improved. Therefore, those changes are considered as a part of the invention which points out the claims.

Although the preferred embodiments have been described here, it is apparent to those skilled in the art that various of modifications, complements, replacements and similar changes can be made without departing from the spirit of the invention, therefor those alternations are considered to be in the inventive scope defined by the appended claims.